知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

网络存档数据质量保证策略理论框架研究

◎ 王文玲[1] 曲云鹏[2]

[1] 国家图书馆 北京 100081

[2] 中国科学院文献情报中心 北京 100190

摘要：[目的/意义]数据质量保证工作是网络存档工作中的一项重要工作，其贯穿整个网络存档工作的始终，决定网络资源存档工作的成败。[方法/过程]通过对国内外各保存机构的质量保证策略及方法进行分析、研究和对比，提出数据质量保证的策略理论框架。[结果/结论]该框架以数据为中心，制定一系列的业务标准及工作规范，利用现有软件工具开展全流程的数据质量检查工作，同时以团队建设、运行环境维护及授权获取网站备份作为补充手段，确保获取高质量的存档数据。

关键词：网络资源存档　质量保证　质量检查

分类号：G251

引用格式：王文玲，曲云鹏．网络存档数据质量保证策略理论框架研究 [J/OL]. 知识管理论坛，2018, 3(2): 106-115[引用日期]. http://www.kmf.ac.cn/p/131/.

1 引言

在网络资源存档工作实践中，会遇到很多类型的数据质量问题，包括网站内容文件的缺失、多媒体内容无法展现、版式错乱等。如果对这些数据质量问题不采取严格的质量控制手段，则可能丢失很多重要的信息，导致数据质量偏低，甚至保存任务失败。根据互联网档案馆（Internet Archive，IA）的技术团队 2013 年发布的“网络资源生命周期模型”[1]，质量检查与分析工作处于网络资源存档生命周期的内环，是上一轮采集存储与组织工作和下一轮采集评估与选择工作之间的重要步骤，决定了下一步工作的方向，可以说网络资源存档的数据质量保证工作是影响网络资源存档成败的主要因素之一。

一般来说，外观完整性、交互完备性和数据一致性被视为存档数据质量的三大评价指标。高质量的网络存档数据指在尽量短的时间内，完整采集目标网站中的知识内容，并且完整保存网站的视觉内容和浏览体验。不同的采集机构会根据各自的采集需求、成本预算等制定不同的量化指标来定义本机构适用的网络存档数据质量评价标准。网络存档的数据质量保证工作即指网络存档机构为保证所采集的网络资源达到预设的质量标准而采取的相关措施和方法，包括机器自动执行、人工干预等方式，范围覆盖采集前、采集过程以及采集后整个工作流程[2]。

作者简介：王文玲 (ORCID: 0000-0002-0236-6191)，馆员，硕士，E-mail: wangwl331@163.com; 曲云鹏 (ORCID: 0000-0002-1611-0904)，副研究馆员，博士。

收稿日期：2018-03-19　　发表日期：2018-04-27　　本文责任编辑：刘远颖

知识管理论坛，2018(2):106-115
DOI：10.13266/j.issn.2095-5472.2018.011

由于网络资源存档技术的限制和网络资源的复杂性，对网络资源进行完美的存档是不实际的，国际上多个网络存档机构，如国际互联网保存组织（International Internet Preservation Consortium，IIPC）、美国国会图书馆、法国国家图书馆等，都开展了相应的数据质量保证相关工作，他们从各自的网络资源存档需求出发，对网络资源存档中的数据质量保证问题进行研究，制定符合自身要求的质量保证策略与方法，尝试从不同程度上解决数据质量问题，尽量提高采集数据质量。

本研究的主要目的是通过调研国内外网络资源存档机构的质量保证工作实践，归纳总结并深入分析质量保证的方法和手段，提出具有普适性的数据质量保证策略理论框架。

2 网络存档质量保证相关实践

2014 年美国北德克萨斯大学的 B. R. Ayala 等人在 IIPC 的资助下开展了网络资源存档质量保证实践的调研[3]，该调研主要面向 IIPC 成员机构，也包括一些非 IIPC 成员的机构，调研方式包括文档分析、邮件交流、会议和当面交流等，调研内容覆盖了网络资源存档数据质量问题本身、机构对质量检查的态度、质量保证的手段和方式以及各种质量问题的解决方案等。调查结果显示：绝大多数机构在进行网络资源采集时都会同时开展质量检查相关工作，只有不到 5% 的机构从来不进行质量检查，对采集过程开展全流程质量控制的机构达到 11.1%。由此可见，网络资源存档机构对质量保证工作都非常重视，数据质量被认为是网络资源存档工作中最重要的问题之一。下文将介绍国内外网络资源存档机构的质量保证实践工作。

2.1 国外网络资源存档机构

2.1.1 美国德雷萨尔大学

德雷萨尔大学（Drexel University）开展了面向高等教育资源的网络存档工作[4]，采用 Archive-It 作为保存工具，主要通过工作人员手动检查来完成质量控制，从而保证重要内容的可用性。工作人员使用 Excel 表单来记录种子采集中发生的错误，每次采集完成后，工作人员需记录种子的采集情况，然后检查 Archive-It 自动生成的质量保证报告中的基础性问题，如采集数量是否过大、数据是否已入存储队列、是否遵守蜘蛛协议（robots.txt）等。之后工作人员开始查找种子错误（种子是否将采集误导到其他网站）以及内嵌文档问题（丢失内容或显示失败）。这些基础的质量控制完成之后，工作人员做出必要的修改及改变，执行修补采集或者重新采集确保存档内容能够与原网站内容保持一致。

2.1.2 法国国家图书馆

法国国家图书馆于 2006 年发布的工作报告中，描述了他们在网络资源法定缴存工作中出现的各种质量问题，并分享了所采取的质量控制手段和方法。他们认为机器自动采集的海量数据的质量检查方法应该根据采集数据的体量以及结构来确定[5]：

（1）对于广域采集，数据零乱无序，质量检查的主要任务是对数据进行检查、描绘并验证有效性，从而确保能准确地进行储藏和保存。采取的主要方法是收集采集日志报告并进行分析、检查通用技术环境和软件运行状态、对采集数据进行抽样以验证数据的可抽取性和可访问性。

（2）对限定种子列表数量的重点采集或对特定网站的定期采集，资源相对整齐有序，应当对采集资源开展系统的验证与检查。他们开发了一个工具组件来开展更精炼的自动化检验工作，第一个模块可以去除 URL 中不可见字符，并对 URL 进行查重；第二个模块可以检验 URL 有效性，检验 URL 是否已存档，检测网站是否有蜘蛛协议，分析网站地理定位等；第三个模块可以自动对比种子列表与现有采集日志报告，该功能对采集过程中的质量检查尤为有用。

2.1.3 美国密歇根大学本特利历史图书馆

本特利历史图书馆（Bentley Historical Library, BHL）开展了“大学档案与记录项目”（UARP）和“密歇根历史专题”（MHC）两



个网络存档项目，于 2011 年发布了这两个项目的质量控制指南与规程，后来又进行了多次修改完善。该指南分析了网络存档过程中可能遇到的内容及技术问题，提出了质量控制的详细规程及操作规范 [6]：①确认质量控制的目标，检查 WAS（Web archiving service，Archive-It 的前身）质量控制工具的记录报告；②确认存档过程成功启动、采集过程完成；③核实采集设置的正确性、元数据的准确性；④判断采集过程是否有特别重要的内容丢失（特别重要的内容指对理解网站主要内容或关键功能不可或缺的内容，没有必要特别在意个别图像、音频、视频及文本的缺失，除非这些内容对网站的研究价值非常重要）；⑤通过改变采集设置、联系网站所有人或对网站进行重新采集来解决一些突出的质量问题；⑥详细记录整个质量控制的处理过程。

2.1.4 美国国会图书馆

上述几个网络资源存档机构主要采用人工质量检查手段，辅助以采集软件的质量控制功能或定制开发的简易质量控制工具，美国国会图书馆与互联网档案馆合作，开展了半自动化质量保证的尝试 [3]，尽管在质量保证过程中也需要大量人工操作，但在某些环节已经全部实现自动化。美国国会图书馆基于采集频率的网络资源存档采集流程如下：

（1）预采集。预采集只采集种子的主页，目的是检测种子列表或 SURT 格式 (种子列表附属文件) 是否有问题。若发现问题会实时对种子列表进行调整。

（2）采集。按照既定的采集频率开展采集工作，期间检测到的任何问题会在 24 小时内报告给网络资源存档团队。采集结束后为采集数据生成 CDX 文件（所采集 URL 的索引文件）和 WAT 文件（所有 WARC 文件的元数据文件）。

（3）自动化质量保证工作。首先进行浏览器分析模拟，该过程使用浏览器模拟器 PhantomJS 以及回放软件 Wayback 对较为重要的种子进行回放，对网站页面进行快照并记录响应代码，生成每个页面的报告，抽取报告中丢失的、需要重新采集的文件列表，添加到采集软件中进行补充采集。同时使用 Pig 脚本对 WAT 索引文件进行链接分析，根据链接类型对外链进行分类，查看所有外链对象尤其是内嵌对象（如 CSS 和 JS 文件）是否被采集，使用 Hadoop 工具对比 CDX 文件中所采集资源与外链对象的差异，从而得到未采集到的对象资源，加入补充采集的候采名单。

（4）补充采集。识别质量问题是回放的问题还是采集的问题，并采集需要补充的内容。

（5）人工质量保证过程。数据管理员浏览存档内容，以代理模式进行观察，检查日志报告，查看所需要的内容是否全部采集。

采集团队发现 [3]，自动化质量保证过程大大提高了采集质量，但是回放质量并没有明显提高。

2.2 国内网络资源存档机构

2.2.1 北京大学

北京大学网络与分布式系统研究所早在 2001 年就搭建了一个大规模的 Web 存档系统——Web Infomall[7]，该系统致力于对中国互联网网页进行存档、组织并提供服务。截至 2013 年 9 月，该系统保存网页 85 亿，数据量达到 73TB。相比其他网络存档项目，Web Infomall 的采集策略比较简单，只采集网页中的静态信息进行存储，网页存储采用自行开发的天网存储格式，采取增量方式对网页进行存档。该项目所采集的网页信息，一方面通过通用公共许可免费分发给需要使用的研究机构，另一方面作为数据挖掘研究与应用的语料库，为 4 个衍生数据服务系统提供数据源。在满足项目既定需求与目标的前提下，仅采集静态信息的存档策略大大降低了采集的难度和失败几率，加之结构化天网存储格式具备一定的容错性，使得采集数据的质量能够得到一定的保障。

2.2.2 国家图书馆

作为 IIPC 唯一的中国成员，国家图书馆从 2003 年开始开展网络信息资源采集与保存实验

知识管理论坛，2018(2):106-115
DOI：10.13266/j.issn.2095-5472.2018.011

研究，并于2009年成立了国家图书馆互联网信息资源保存保护中心，通过十几年的摸索与实践，目前已经形成规模化采集，截至2016年底，网络导航和网络资源采集总量达到114.73TB[8]。国家图书馆的网络存档有两种类型，一种是网站采集，主要针对政府网站、组织机构等，采用广度优先采集策略；另一种是定题采集，主要针对重大事件，如"党的十九大"等，采用深度优先采集策略。不论是全域采集还是专题采集，都专门制定了详细的资源采选原则及标准以及种子重要性排序原则。在数据检查方面，充分利用工具软件对采集页面进行回放检查，建立相应的检查机制，制定数据检查工作流程及操作规范，对检查方式和抽样率等进行规定。

2.3 质量保证实践对比分析

德雷萨尔大学的质量控制工作发生在抓取工作之后，手段是采集日志及采集软件质量报告分析，方式为人工手动检查。法国国家图书馆对广域采集和专题采集采用不同的质量保证策略，广域采集采取日志分析、抽样检查和软硬件运行状态检查，专题采集使用定制开发工具进行半自动化质量控制。本特利历史图书馆从策略层面制定了详尽的质量保证工作流程规范，尽可能减少因个人因素导致的质量问题。美国国会图书馆将质量保证工作扩展到采集工作的各个环节，加入预采集环节，对抓取过程进行监控，使用软件回放并自动记录质量问题，不仅使采集工作流程得到优化，整个质量控制过程也实现了相当高程度的自动化。北京大学主要考虑网络资源时效性的特点，希望尽快将网络资源采集到本地进行保存，对数据质量要求并不高，仅对采集结果进行格式检查。国家图书馆目前着眼于发动国内有能力的公共图书馆参与网络资源的联合建设工作，因此制定了尽可能详细的采集策略以及严格的质量保证工作规范，质量检查仍以人工分析日志以及回放检查为主，以确保采集数据质量的均一性。如表1所示：

表1 质量保证实践对比

机构名称	流程、规范	保障环节	检查手段	自动化程度	时效保障
德雷萨尔大学	—	采集后	日志分析、质量报告分析	人工	一般
法国国家图书馆	—	采集前、后	日志分析、抽样检查	半自动化	一般
本特利历史图书馆	非常详尽	采集后	质量报告分析	人工	一般
美国国会图书馆	非常详尽	采集前、中、后	自动回放、日志分析	高自动化	一般
北京大学	—	采集后	格式检查	人工	高
国家图书馆	详尽	采集后	日志监控、回放检查	人工	一般

注：表中"—"表示本文作者所掌握的材料未提及，不代表该机构未采取相关措施

综上可以看出，由于在采集需求、成本预算、时效性等方面要求各异，每个存档机构在质量保证工作中使用的方法手段以及投入的时间精力也各不相同，各机构都根据各自项目特点选择了适用于本项目的质量保证措施和手段。显而易见的是，质量保证措施和手段越复杂精细，所得到的数据质量越高。

3 网络资源存档数据质量保证的策略框架体系

在不考虑项目具体采集需求及项目成本的前提下，为追求尽可能高的存档数据质量，本文提出了网络资源存档数据质量保证的策略理论框架（见图1），该框架以数据为中心，制定一系列的业务标准及工作规范，利用现有软件

工具对采集过程开展全流程的数据质量检查工作，同时以团队建设、环境维护及授权获取网站备份作为补充手段，确保获取高质量的存档数据。

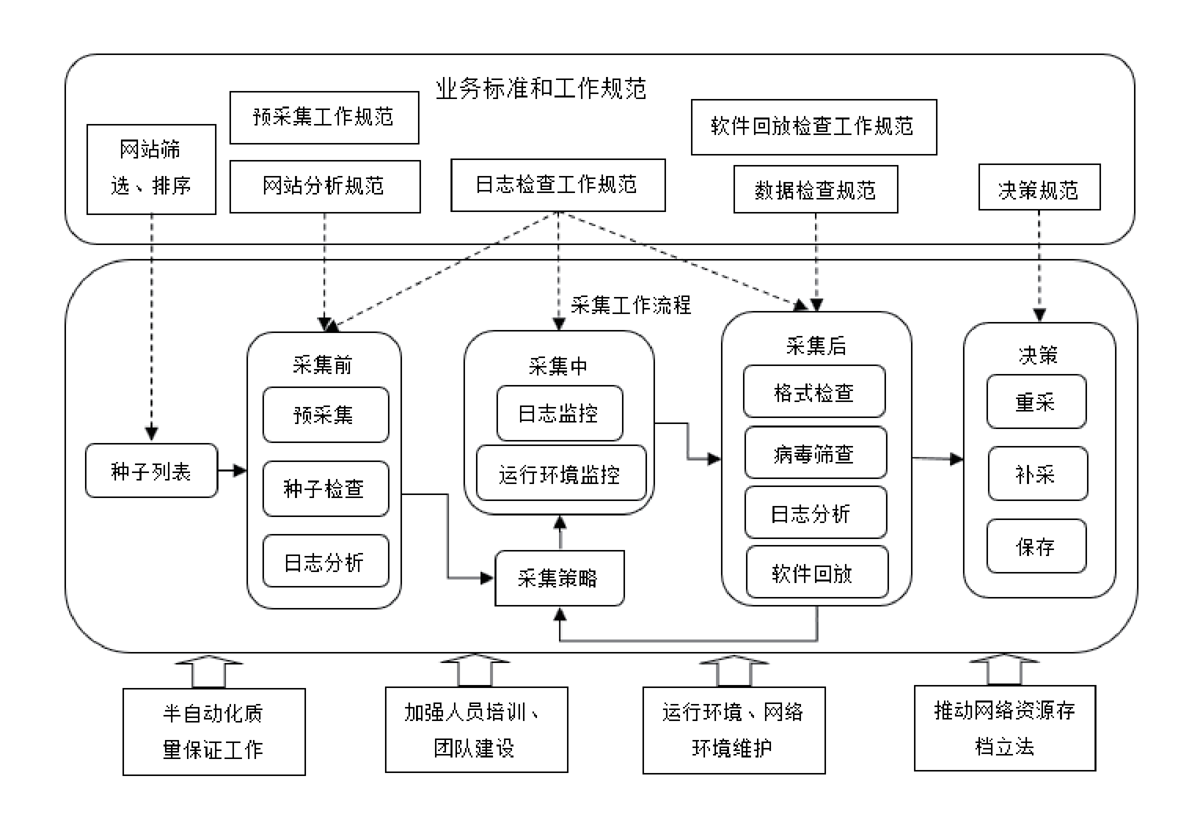


图 1 网络资源存档数据质量保证策略框架

3.1 制定严格的采集业务标准和工作规范

从上文调研情况可知，网络存档质量保证工作以人工操作为主，而质量控制专家背景知识、技术水平及工作熟练程度各不相同，为避免人为因素导致的数据质量问题，应当为网络存档工作制定统一的业务标准和严格的工作规范。推荐制定的标准规范包括：

（1）数据质量标准。前文提到的高质量网络存档数据概念，只是一个理想化的定性描述，存档机构应当在完整保存知识性内容、完整保存视觉内容和浏览体验、尽快地完成采集任务这三者之间进行平衡，根据采集任务的目的和需求为数据质量制定量化的便于操作的标准，这是一切质量保证工作的基础。

（2）数据格式标准及元数据规范。常见的网络存档数据格式有 WARC、ARC 和 KW 等，其中 WARC 是国际标准兼国家标准，是网络存档领域首选的数据标准，ARC 和 KW 是行业标准，正在逐步被 WARC 所取代。制定存档网站对象元数据规范便于将来对网站对象和存档数据进行管理，内容应该包括网站题名、主要内容标签、采集时间、容量、URL 数量、地理位置等，其中地理位置和主要内容标签可以用于筛选网站是否符合采集要求。

（3）软件使用标准。网络存档工作需要使用的软件工具包括采集软件、分布式存储软件、杀毒软件、回放软件等。应制定相应的软件使用标准，明确各类软件工具的选择范围、版本以及标准配置等。例如，目前最常用的采集软件是 Heritrix 和 Wget，常用的面向用户桌面的工具是 WarcCreate 软件，在业务工作中可限定只允许使用 Heritrix 软件进行采集，并对所使

知识管理论坛，2018(2):106-115
DOI：10.13266/j.issn.2095-5472.2018.011

用的规则进行严格限制，以确保生成数据的一致性。

（4）种子筛选、排序标准。无论是国域采集还是专题采集，可以参考一些知名数据统计排名进行种子筛选，例如 Alexa 排名等，也可以参考一些现有的、较为权威的网站列表等。对种子网站的选择设定一个标准，可以保证原网站的数据有较好的质量。种子筛选完成之后一般会依照某些特征，对符合要求的种子进行优先级排序，并根据需要设定每个种子网站的采集时间、采集频率、采集范围等。制定种子排序标准可以更有针对性、更有序地开展网络资源存档。

（5）爬虫默认配置。深度采集和广度采集是最常使用的两种采集方式。采集团队应该根据自己业务的需要，提供这两种采集方式的爬虫默认配置，然后根据具体的种子采集需要，对爬虫配置进行尽可能小的修改。这样能尽量减少主观原因导致的爬虫参数配置错误，从而提高采集的数据质量。

（6）各类操作规范。网络存档是一项人工干预程度比较高的工作，鉴于人工操作的随意性，为每一个程式化的步骤制订便于执行的工作流程及规范将大大提高每一步操作的规范性，降低错误发生的几率，例如：预采集工作流程及规范、日志检查工作流程及规范、病毒筛查工作流程及规范、软件回放质量检查工作流程及规范等。

3.2 开展全面的数据检查

开展直接的数据质量检查是保证数据质量最有效的工作之一，是质量保证工作的核心，应贯穿到网络资源存档的整个工作流程当中。根据质量检查工作开展的时间，数据检查可以分为采集前检查、采集中检查和采集后检查，这 3 个阶段质量检查工作的目的和内容均不相同。

3.2.1 采集前检查

采集前对目标网站的结构和内容进行预采集和分析是成功采集的重要前提，通过分析可以及时调整采集策略，从而避免多域名、外链引用等导致的质量问题，也可以帮助质量控制专家明确网站内容是否符合采集要求、所需内容是否符合网站蜘蛛协议等。预采集和分析的目的是确定目标网站的采集策略，从而提高采集数据质量。预采集一般采用通用采集策略和配置进行采集，采集的内容通常不会写入本地文件，只获取爬虫采集的日志，从而对日志进行分析。

在日志分析过程中，质量控制专家需要识别不完整、不准确或不成功的网络采集结果，判定不合标准的采集并找出其内在原因或问题。这一步骤可能需要确认爬行设置，复审爬行报告和日志，检查目标网站的内容、布局、特征和源代码，记录可能阻止准确可靠并精确捕捉网站的任何技术限制、爬虫协议的除外条款或其他问题。通过预分析，需要确定本次采集的深度、广度和采集频率；确定是否存在大量脚本生成的资源而导致采集失败；确定是否存在爬虫陷阱；确定网站中所需要采集的内容都分布在哪些服务器；确定采集规则应如何设定；确定种子列表是否完备以及是否出现错误等。明确以上内容即明确了采集策略，可以开始进行正式的采集。

3.2.2 采集中检查

采集中检查的主要目的是确保爬虫正常运行，顺利完成采集任务。采集中的检查任务包括：①实时监控爬虫运行是否正常，是否存在内存溢出等技术问题？是否陷入爬虫陷阱？网络情况是否正常？②定期检查采集日志，判断需要采集的内容是否正确采集？采集的资源是否所需？是否出现预采集没有发现的问题？爬虫设置是否合理？是否需要修改？通过这些检查，及时解决爬行过程中发生的问题，可以保证爬虫顺利完成采集任务。

3.2.3 采集后检查

采集后的质量检查工作主要内容是：对数据进行校验，判断是否满足既定格式、查杀病毒；检查存档文件是否能展现网站原貌，采用的方法有日志分析和软件回放。

(1) 数据校验。数据检查的第一步是查看数据是否符合既定的数据标准，对数据文件进行置标校验，确保文件的完整性。通常网络资源存档机构都会采用国际标准 WARC 格式来作为既定数据标准，此步骤可以通过 JHove2、WARC Tools 等工具来完成。

(2) 查杀病毒。查杀病毒不是数据检查的必备步骤，笔者认为，网页存在病毒是当代网络的一种特征，其原因可能是网站被黑客攻击，也有可能是网站管理员的误操作，也有可能是杀毒软件的误报。从网络资源存档的业务工作角度讲，数据质量有两个标准：一是满足预定的目标，二是反映客观的实际。如果存档的目的是为了服务，那么采集到的资源一定不能有病毒；如果存档是为了保存，那么对于病毒的存在应该抱容忍的态度。

病毒的查杀可以使用常用的杀毒软件（如卡巴斯基、Avira 等）来进行，操作比较简单。病毒查杀的关键在于对病毒的处理，有的杀毒软件会把病毒文件从采集生成的 WARC 文件中移除；有的杀毒软件是直接把整个 WARC 文件删除，这样必然导致采集数据的丢失；还有的杀毒软件对脚本比较敏感，尽管这些脚本是无害的，会导致误杀。因此，在进行杀毒之前，必须对所使用的杀毒软件有所了解，并选择正确的杀毒软件进行。

(3) 日志分析。经验丰富的质量控制专家可以在日志分析过程中发现自动跳转、引用外链等技术问题，同时也能根据采集到的文件数量和采集任务持续的时间发现相关问题。在采集日志中需要特别注意服务器响应错误及超时错误，若对方服务器或网络出现了问题，导致无法访问，网络爬虫通常不会持续地尝试访问目标资源，而是在多次尝试失败之后在日志中记录，然后跳过该资源，这样便会造成信息漏采，因此在检查日志时需要重点监控这类信息，对漏采的资源进行分析，必要时加入补采清单。

(4) 软件回放。使用专门的回放软件对存档内容进行回放，再通过人工点击和查看的方式，来确认网站内容是否完整、链接是否有效、交互性是否完备。由于软件回放需要全部由人工来操作，对于海量采集这种方式变得很难执行，只能进行抽样检查，作为日志分析的补充手段。

3.3 开展半自动化质量保证工作

为提高网络存档质量保证工作的自动化程度，减少人工参与度与工作量，越来越多的网络资源存档软件工具开始集成质量保证的功能模块，也有不少专门的质量保证辅助工具出现。若能充分利用这些功能模块和工具，质量控制专家则能事半功倍地完成质量保证工作。

3.3.1 爬虫软件

Heritrix 是网络资源存档中使用最广泛的爬虫软件 [9]，它为采集前、采集中到采集后的质量控制都提供了良好的支持。Heritrix 拥有非常强大的任务配置功能，它提供了数十条不同颗粒度的采集规则，包括采集范围、采集协议、抽取内容类型、文件输出格式等。同时为符合颗粒度规则的 URL 提供正则表达式过滤、目录深度过滤、跳转次数过滤、SURT 过滤等进一步的过滤选择。通过这些规则的灵活组合，任务管理员可以制定与采集目标高度吻合的采集策略，从而为各种复杂程度不同的采集任务提供高质量的采集结果。

Heritrix 提供监控采集任务的控制台，可实时查看当前任务进度、采集速度、运行时间、采集线程、队列情况等信息。若任务运行过程当中遇到访问超时、任务无法启动、网页解析错误等问题，控制台还会发出警报提醒，以便任务管理员及时查看并排除故障。

此外 Heritrix 还为采集任务提供详细的报告，里面记录本次任务所采集的资源数量、类型、容量、所属类型等信息。任务管理员可以通过这些信息，分析每个域名的采集情况，并判断采集失败的原因，以便下次采集时根据失败的原因重新设定新的采集规则，以保证采集内容的完整性。

3.3.2 采集软件包

Web Curator Tools（WCT）是一款开源的

DOI：10.13266/j.issn.2095-5472.2018.011

网络资源存档工具软件，功能包括采集授权管理、采集任务安排、质量检查、采集数据验证、元数据管理等。WCT利用Heritrix作为采集爬虫，同样能提供采集前、采集中到采集后的全流程质量控制。WCT在质量控制方面最为突出的是其专门提供一个图形化的质量检查工具，用户操作界面以种子URL为根，将所采集到的资源URL以树形结构的方式进行展示，同时显示每个URL的部分统计信息，包括URL总数量、成功采集的数量、采集失败的数量、对象容量大小等。质量控制专家可以直接对树形结构中的"枝叶"进行修剪，删除不需要的资源内容，也可以在结果中导入遗漏的URL或遗漏的单个文件等。修改完毕进行保存时，WCT会自动对存档资源进行相应的修改并更新。

NetArchiveSuite是一套完整的网络资源存档软件包，由丹麦皇家图书馆和洲际大学图书馆联合开发[10]，主要用于网络资源存档工作规划、采集任务安排、网络资源采集等。NetArchiveSuite提供了一个专门用于回放的质量检查工具ViewProxy，功能包括：①结合浏览器模拟器对网络存档资源进行回放浏览；②收集采集过程中丢失的URL，加入补充采集队列；③直接对丢失的URL进行再次采集。

3.3.3 辅助工具

Monitrix是一个专门为Heritrix 3设计的前端监控、分析软件，主要功能包括：①实时监控Heritrix 3的任务运行情况，生成可视化的图形，显示各种统计信息；②根据采集日志，生成采集时间线，显示单位时间的统计信息，包括数据量、URL数量、发现的新主机数量、完成采集的主机数量；③浏览详细的统计信息，包括主机数量、URL数量、日志中的警告数量；④针对单个主机进行信息分析，包括第一次访问时间、最后一次访问时间、HTTP相应代码饼图、病毒检查饼图、MIME资源类型饼图和子域的列表等。

3.3.4 专门质量检查软件

Jhove2[11]是知名的开源格式验证软件，在长期保存领域得到了广泛的应用，从2.1.0版本开始支持WARC标准文档的分析和验证。JWAT（Java web archive toolkits）[12]为用户提供了图形化界面，不仅可以读取和验证WARC/ARC文档，还提供了写入功能，方便用户及时纠错。Warc tools[13]是由IIPC资助开发的WARC文档处理工具，提供多个脚本来实现处理WARC文档的功能，如WARC文档的验证、生成自动摘要、ARC文档转换为WARC文档等。

Wayback machine[14]是互联网档案馆开发的WARC/ARC文档索引和回放软件，它支持对WARC/ARC文档中的URL进行索引和回放，并提供用户检索界面。OpenWayback[15]是Wayback machine的Java版本，由国际互联网保存组织主导开发，实现了Wayback machine的大部分功能，是目前主流的回放软件。

3.4 其他策略

3.4.1 加强网络资源存档团队建设

网络资源存档数据质量保证工作目前主要通过人工完成，质量控制专家的专业能力直接影响质量保证工作的效果。作为一名合格的质量控制专家，应熟练掌握互联网相关知识，包括互联网数据传输技术、网站开发技术、网络硬件相关知识，除此之外还需具备较强的数学能力、逻辑推理能力和编程能力等。面对互联网技术突飞猛进的发展速度，应当加强对网络资源存档团队的培训，提升其专业能力，使其成为数据质量保证工作的人才保障。

3.4.2 做好运行环境及网络环境维护

网络存档工作的成败不仅取决于各种网络采集软件及工具的功能，也依赖于软件工具运行的硬件环境及网络环境，维护良好的软硬件运行环境及网络环境是保证高质量网络存档工作的前提。网络采集团队应制定严密的服务器及硬件管理规定，运用各种监控网络硬件的设备及软件，定期对服务器软硬件运行环境进行检查，为网络存档工作提供良好的软硬件运行环境及网络环境。

3.4.3 直接获取网站资源备份

网络爬虫是一种有缺陷的网络资源存档技

术，它模拟人类浏览网页时的情形，但又因缺乏智能性不能完整模拟，因此这种方法永远不能完美呈现原始网站面貌。保存机构若与网络资源所有者进行合作，在解决知识产权等相关问题的前提下，直接从提供商处获取网站及资源的数据备份，包括后台数据库、嵌入式资源以及动态脚本等，大概就是网络资源存档的“终极”解决方案了。虽然这种方式目前来看操作性不强，但是一条值得探索的道路，可以在小范围内进行尝试。

3.5 小结与建议

本文通过调研分析国内外网络存档机构在数据质量保证方面的措施和方法，提出了通用的具有普适性的存档数据质量保证策略理论框架。本框架的提出不基于任何具体的采集项目需求，也不考虑质量保证措施所耗费的人力成本和物力成本，是一个通用的可供存档机构参考选择的理论框架。在外观完整性、交互完备性和数据一致性三大质量评价指标中，本框架更注重外观完整性和交互完备性的保障，对数据一致性也就是采集时效性考虑较少。D. Denev[16] 等人提出的 SHARC 框架，采用一系列注重质量的采集时间策略，增加对频繁变化网页的采集频率来保证数据的一致性。存档机构可根据各项目采集目标及采集需求，充分考虑项目成本预算，对框架中的质量保证具体方法和手段进行有针对性的选择。

4 结语

目前的网络存档工作面临的两大难题为知识产权问题和爬虫技术问题。我国网络资源存档呈缴立法工作尚处于空缺状态，应当从网络文化保护的角度出发，推动网络资源呈缴的相关立法工作，这样保存机构就可以突破蜘蛛协议等技术性限制，尽可能完整地保存网站内容。在法律问题解决之前，利用网络爬虫进行采集仍然是主要的手段之一，在未来的工作中，网络存档机构应当努力增强网络爬虫的采集能力，解决富应用封装网络资源的采集问题，从而提高采集的质量。

参考文献：

[1] BRAGG M, HANNA K. The Web Archiving Life Cycle Model[EB/OL]. [2018-03-12]. https://archive-it.org/static/files/archiveit_life_cycle_model.pdf.

[2] 王文玲 , 曲云鹏 . 网络资源存档数据质量问题初探 [J]. 数字图书馆论坛 , 2018(4):8-13.

[3] AYALA B R, PHILLIPS M, KO L. Current quality assurance practices in Web archiving [EB/OL]. [2018-02-05].https://digital.library.unt.edu/ark:/67531/metadc333026/m2/1/high_res_d/QA_in_WebArchiving.pdf.

[4] ANTRACOLI A, DUCKWORTH S, SILVA J. Capture all the URLs: first steps in Web archiving [EB/OL].[2018-03-01]. http://palrap.pitt.edu/ojs/index.php/palrap/article/view/67/370.

[5] ILLIEN G. Sketching and checking quality for Web archives: a first stage report from BnF[EB/OL]. [2016-05-05]. http://bibnum.bnf.fr/conservation/bnf-qualityforwebarchives-feb06.pdf.

[6] SHALLCROSS M. Quality assurance for the Bentley Historical Library Web archives: guidelines and procedures[EB/OL]. [2018-03-01]. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/94162/BHL_WebArchivesQA-v3-20130909.pdf.

[7] 闫宏飞 , 黄连恩 , 谢正茂 , 等 .Web Infomall：一个大规模的 Web 存档系统 [C]//. 网络资源采集与数字资源长期保存学术研讨会论文集 . 北京 : 国家图书馆出版社 , 2013.

[8] 国家图书馆 . 国家图书馆 2017 年年鉴 [EB/OL]. [2018-03-12].http://www.nlc.cn/dsb_footer/gygt/ndbg/nj2017/201712/P020171220578252136424.pdf.

[9] Heritrix[EB/OL]. [2018-03-12]. https://webarchive.jira.com/wiki/spaces/Heritrix/overview.

[10] NetArchiveSuite[EB/OL].[2018-03-12]. https://sbforge.org/display/NAS/NetarchiveSuite.

[11] JHOVE2[EB/OL]. [2018-03-12]. https://bitbucket.org/jhove2/main/wiki/Home.

[12] CLARKE N.Java Web archive toolkit[EB/OL]. [2018-03-18]. https://sbforge.org/display/JWAT/Overview.

[13] Hanzo.WARC Tools project[EB/OL]. [2018-03-18]. http://netpreserve.org/projects/warc-tools-project/.

[14] Wayback machine[EB/OL]. [2018-03-18]. http://wayback.archive-it.org/.

知识管理论坛，2018(2):106−115
DOI：10.13266/j.issn.2095−5472.2018.011

[15] OpenWayback [EB/OL]. [2018-03-18]. https://github.com/iipc/openwayback/wiki.
[16] DENEV D, MAZEIKA A, SPANIOL M. The SHARC Framework for data quality in Web archiving[EB/OL]. [2018-03-12]. https://domino.mpi-inf.mpg.de/intranet/ag5/ag5publ.nsf/AuthorEditorIndividualView/0de8d19ced5a8ae7c1257849005270a3/$FILE/sharc-vldbj.pdf.

作者贡献说明：
王文玲：负责资料收集、分析和论文撰写；
曲云鹏：提出论文写作思路，修订完善论文。

Research on the Theoretical Framework of Web Archiving Data Quality Assurance Strategies

Wang Wenling[1] Qu Yunpeng[2]
[1]National Library of China, Beijing 100081
[2]National Science Library, Chinese Academy of Science, Beijing 100190

Abstract: [Purpose/significance] Quality assurance is one of the most important procedures in web archiving, it runs throughout the whole web archiving work and affects the success odds of web archiving work. **[Method/process]** In this article, we made an analysis and comparative study for the quality assurance strategies of domestic and foreign web archiving organizations, and proposed a strategic theoretical framework for data quality assurance. **[Result/conclusion]** The framework in this article is a data-centered design, it includes a series of criteria and operating specifications, carries out data quality inspection throughout the collecting procedure by using semi-automatic auxiliary tools. Meanwhile, to ensure access to high quality archive data, the framework also takes team building, running environment maintenance and authorized backup to the websites as supplementary means.

Keywords: web archiving quality assurance quality inspection